Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia - Social and Behavioral Sciences 47 (2012) 1077 – 1082
Procedia
Social and Behavioral Sciences

CY-ICER 2012

Use of the program Winsteps for analyzing test tasks and test persons

Prextová Tatiana

University of Ostrava, Pedagogical Faculty, Department of Information and Communication Technologies, Fráni Šrámka 3,Ostrava 709 00, Czech Republic

Abstract

Didactic test is a useful and effective instrument to inform teachers about the knowledge, the skills and the knacks of the individual pupils. Didactic test has the several basic advantages: the testing of a wide range of curriculum, the evaluation of a large number of pupils, enabling an objective comparison of the level of the knowledge; it is easy to use, gives us the effective evaluation (implementation in the short date), resulting in a higher reliability. The quality of the created didactic test depends on suitable created test items. The combination unsuitable test tasks generates a test, that is not suitable for testing, it does not measure what would be measured, it is small reliable and don't report to us any new information on the test person. In making the suitable test items, it is important to concentrate on their difficulty. We can find the difficulty of items according to how many pupils answer correct this item, let us say incorrect. Too easy or too hard items don't give us any information about the differences between the test pupils. Determining the difficulty of test items, their comparison, the determining the reliability of tasks and people we can realize with the program Winsteps. The interesting is also the demonstration that in which persons and which items have been surprising answers. The program offers us the possibility of the results in a tabular form or a graphical form or in the form of a summary graph.

Keywords: Program Winsteps, reliability, difficulty, didactic test;

1. Introduction

The quickest way of the verification of the pupils' knowledge, the revision their acquired knowledge is a didactic test. In the didactic test pupils work in a relatively equal conditions, by test is possible to test not only a wide spectrum of subject matter, but also a large quantum of the pupils, an evaluation of pupils' knowledge is implemented in a shorter time.

The quality didactic test should verify several basic characteristics. The most important is reliability of the test. Reliability of the didactic test guarantees us accuracy how test measures and detects pupils' knowledge (Lapitka, 1990). These indicators affect reliability of the test:

- number of tasks in the test
- difficulty of tasks
- number of test pupils (Burjan, 2005)

The other important characteristic of the didactic test is difficulty. Tasks in the order of difficulty show us, which the assignments are difficult for pupils, and on the other side, less difficult.

How can these and other properties determined using Winsteps, we decided to show a sample of 2554 persons tested fifth year of primary school who responded to the 12 tasks in mathematics.

 doi:10.1016/j.sbspro.2012.06.781

2. Persons and reliability of items

There we have data on persons and items are presented in an Excel table. After loading the program Winsteps we see introductory table showing the reliability of persons and items.

Osoba	2554 INPUT		2514 MEASURED			INFIT		OUTFIT	
	TOTAL		COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	7.8		11.2	1.42	.95	.98	.0	.97	.2
S.D.	2.3		1.1	1.39	.30	.37	.9	.92	.7
REAL RMSE	.99	TRUE SD	.97	SEPARATION	.98	Osoba	RELIABILITY	.49	
Položka	12 INPUT		12 MEASURED			INFIT		OUTFIT	
	TOTAL		COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	1634.3		2340.3	.00	.07	.99	-.5	.99	-.6
S.D.	485.8		242.8	1.52	.03	.12	3.8	.29	3.7
REAL RMSE	.07	TRUE SD	1.52	SEPARATION	20.71	Položk	RELIABILITY	1.00	

Figure 1 – Table of reliability of persons and items

The red highlighted areas have reliability test people. This is a lower limit value of reliability. If the reliability value of <0.8, it is necessary to insert more test items. We see that in this case, the number of test items 12 is insufficient.

The green areas have indicated reliability of the test items. Again, this is the lower limit value of reliability. If the reliability value of <0.8, it is necessary to increase the research sample.

3. Difficulty of items

The next generation get detailed table showing the difficulty of test items.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Položka
3	742	2435	2.56	.05	1.35	9.9	1.88	9.9	.30	.53	69.8	77.7	3704
12	942	2284	1.88	.05	.91	-4.3	.92	-2.2	.58	.53	77.8	72.9	3740
8	1233	2329	1.24	.05	.93	-3.4	.94	-1.8	.56	.52	73.0	70.7	3729
4	1523	2427	.70	.05	.91	-4.5	.85	-4.9	.55	.49	75.0	71.7	3721
9	1714	2500	.36	.05	1.02	1.0	1.02	.6	.46	.47	72.9	73.8	3732
7	1728	2475	.29	.05	1.08	3.4	1.10	2.5	.42	.47	71.2	74.4	3726
11	1961	2505	-.29	.05	.95	-1.9	.88	-2.3	.46	.42	81.0	79.5	3738
2	1488	1814	-.57	.07	.89	-2.9	.75	-3.6	.47	.40	84.0	82.5	3698
1	1512	1824	-.65	.07	.90	-2.5	.77	-3.1	.46	.39	84.3	83.3	3689
6	2134	2495	-.90	.06	1.02	.5	1.03	.4	.36	.37	85.5	85.4	3723
5	2170	2493	-1.05	.07	.94	-1.3	.85	-1.8	.39	.36	87.0	86.8	3722
10	2464	2502	-3.57	.17	.98	-.1	.88	-.5	.17	.15	98.4	98.4	3737
MEAN	1634.3	2340.3	.00	.07	.99	-.5	.99	-.6			80.0	79.7	
S.D.	485.8	242.8	1.52	.03	.12	3.8	.29	3.7			8.0	7.7	

Figure 2 – Table of difficulty of items

On the right side of the table entries are arranged from easiest to hardest (top down). In the heaviest item (red areas), we see that the complete number of test subjects (2435, 119 persons did not meet entry), 742 of them correspond to the item correctly. The value of the difficulty of this item is 2.56.

For the lightest line (green area), we see that the total number of test subjects (2502, 52 persons did not match the item), 2464 of which corresponded to the item correctly. Value of item difficulty is -3.57.

4. Latent level of persons

The following table shows us the skill level of individual test subjects.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Osoba
757	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	10944 4 6 1 10 2010 0 1174 929 14066 1
769	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	11083 4 6 1 10 2010 0 237 973 14763 1
793	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	11384 4 6 1 10 2010 0 1174 931 14111 1
806	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	11622 4 6 1 10 2010 0 202 608 9028 1
867	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	12326 4 6 1 10 2010 0 488 1416 22078 1
913	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	13009 4 6 1 10 2010 0 1122 1330 20356 1
943	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	13534 4 6 1 10 2010 0 4032 1446 22509 1
950	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	13604 4 6 1 10 2010 0 1140 1061 17227 1
988	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	13998 4 6 1 10 2010 0 99 1427 22208 1
990	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	14020 4 6 1 10 2010 0 3067 1598 25057 1
1013	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	14464 4 6 1 10 2010 0 3255 719 10236 1
1032	12	12	4.54	1.89	MAXIMUM MEASURE				.00	.00	100.0	100.0	14728 4 6 1 10 2010 0 1283 653 11941 1
747	2	12	-2.18	.93	2.09	1.5	3.38	1.6	-.09	.50	75.0	88.3	10841 4 6 1 10 2010 0 2114 651 11640 1
911	2	12	-2.18	.93	2.11	1.6	3.41	1.6	-.10	.50	75.0	88.3	12997 4 6 1 10 2010 0 239 1478 23411 1
1395	2	12	-2.18	.93	.76	-.2	.51	.0	.62	.50	91.7	88.3	19042 4 6 1 10 2010 0 1131 1700 26901 1
1485	2	12	-2.18	.93	2.26	1.7	3.98	1.8	-.22	.50	75.0	88.3	20616 4 6 1 10 2010 0 569 1692 24867 1
1522	2	12	-2.18	.93	.84	.0	.68	.2	.57	.50	91.7	88.3	20925 4 6 1 10 2010 0 3390 401 5293 1
1778	2	12	-2.18	.93	.93	.1	1.12	.5	.48	.50	91.7	88.3	24718 4 6 1 10 2010 0 1212 1999 32151 1
2056	2	12	-2.18	.93	.78	-.2	.54	.0	.61	.50	91.7	88.3	28494 4 6 1 10 2010 0 1841 2193 34891 1
236	1	10	-3.09	1.30	.30	-.8	.10	-.6	.75	.50	100.0	92.4	2724 4 6 1 10 2010 0 3116 753 10716 1
453	1	10	-3.09	1.30	.30	-.8	.10	-.6	.75	.50	100.0	92.4	5786 4 6 1 10 2010 0 2649 696 10367 1
2365	1	12	-3.32	1.24	.33	-.8	.10	-.7	.71	.44	100.0	92.7	32097 4 6 1 10 2010 0 3334 2271 36439 1
MEAN	7.8	11.2	1.42	.90	.98	.0	.97	.2			79.9	79.7	
S.D.	2.3	1.1	1.39	.30	.37	.9	.92	.7			12.9	6.4	

Figure 3 – Table of skill level

On the right side of the table are ranked according to number of persons tested correct answers in the test. In red we see the person of a total of 12 points gained 12 points. The value of her ability, skill is 4.54.

The green area is the person of the total number of items (12) correctly answered an item. Its ability, skill has a value of -3.32.

5. Graphical representation of difficulty of items

Difficulty of test items can be represented only by a summary table, but also through various types of graphs: Bubble Chart or Graph - Category Probability Curves.

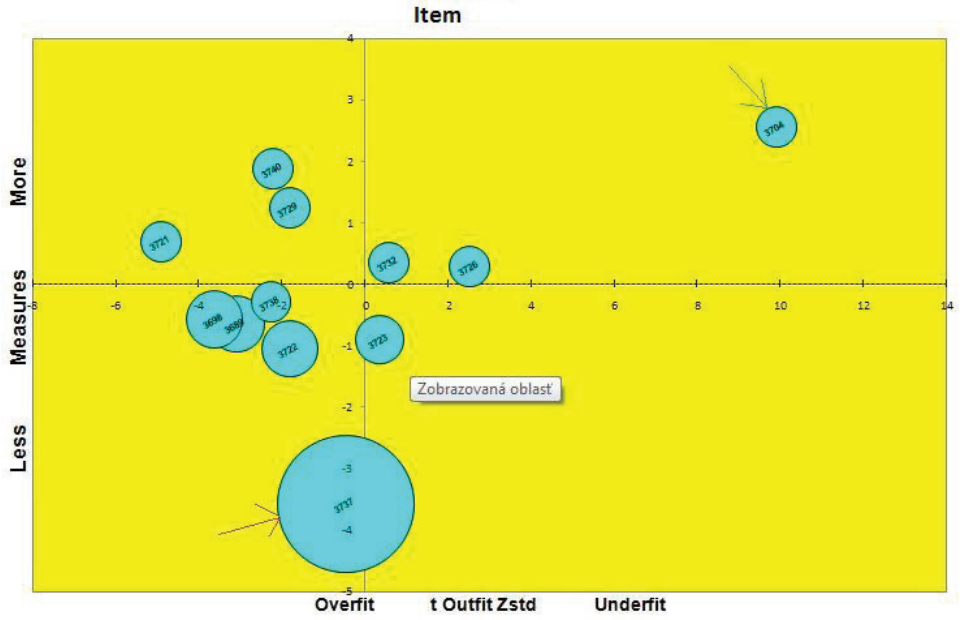


Figure 4 – Bubble chart

On the Bubble chart, we see the distribution of individual test items. A green arrow indicates an the most difficult item in the test. Red arrow indicates the item unobtrusive.

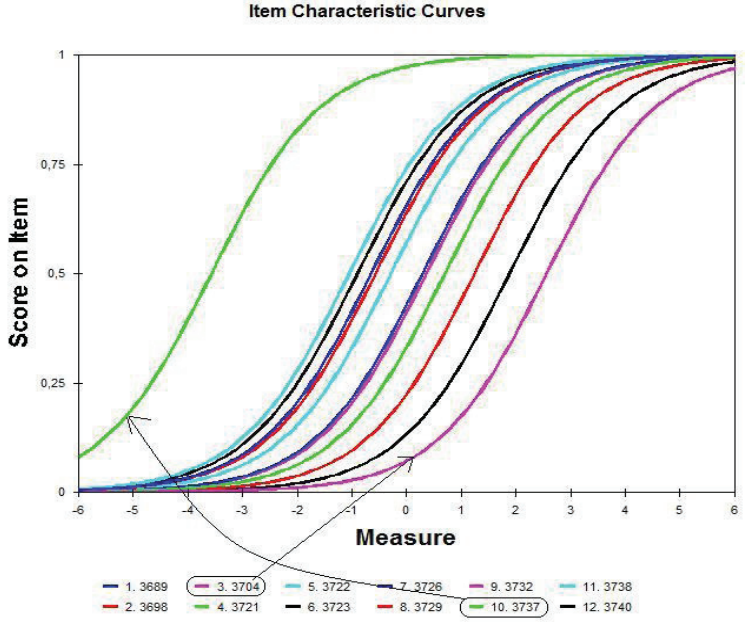
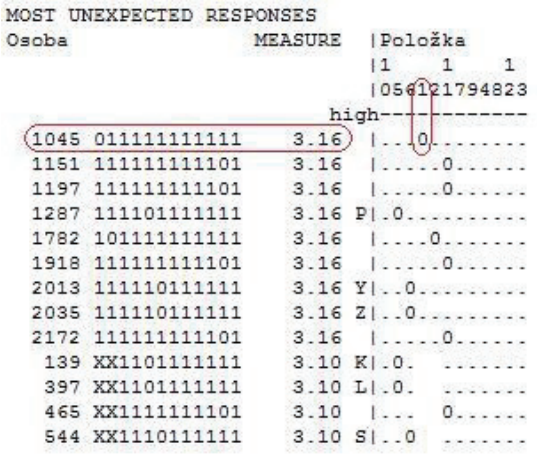


Figure 5 – Category probability curves

On this chart we can see again ranked the items according to difficulty. Left Arrow to us indicates an item that has the lowest difficulty (item 3737), right arrow indicates the item to us, which is the highest difficulty (item 3704).

6. Extreme data tables

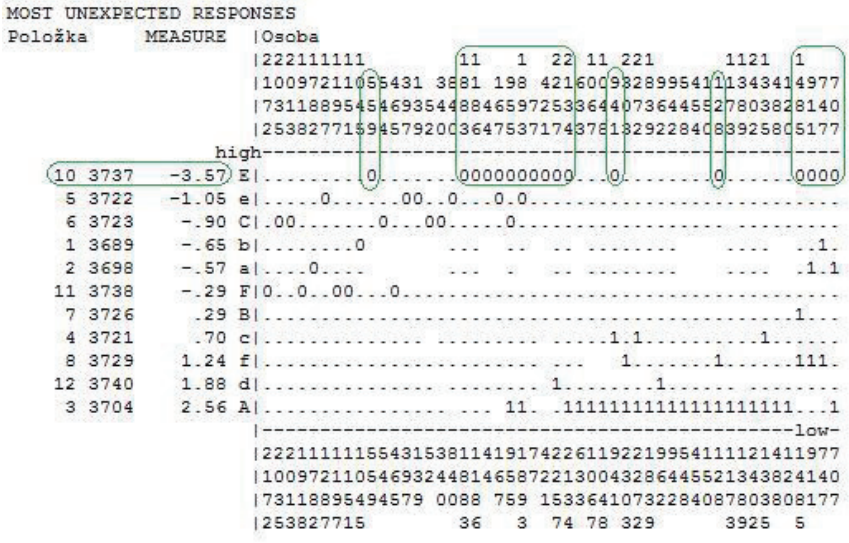
The following tables show us the extreme data items, as well as test subjects.



```
MOST UNEXPECTED RESPONSES
Osoba                MEASURE  |Položka
                              |1    1    1
                              |0561217948 23
                            high----------
1045 011111111111     3.16   |...0.......
1151 111111111101     3.16   |.....0......
1197 111111111101     3.16   |.....0......
1287 111101111111     3.16 P|.0..........
1782 101111111111     3.16   |....0.......
1918 111111111101     3.16   |.....0......
2013 111110111111     3.16 Y|..0.........
2035 111110111111     3.16 Z|..0.........
2172 111111111101     3.16   |.....0......
 139 XX1101111111     3.10 K|.0.  .......
 397 XX1101111111     3.10 L|.0.  .......
 465 XX1111111101     3.10   |...  0......
 544 XX1110111111     3.10 S|..0  .......
```

Figure 6 – Table of extreme data

In this table, we see people who have answers to some of the items have been surprising. The red area (horizontal) is indicated by the person (1045), followed by the individual responses to test items, the value of 3.16 represents the skill level. The red area (vertical) shows that the answer to item 1 is surprising ("." are the answers surprise).



```
MOST UNEXPECTED RESPONSES
Položka     MEASURE  |Osoba
                     |222111111        11   1  22 11 221      1121  1
                     |100972110554 31 3881 198 421600932899541113434 14977
                     |731188954546935448846597253364407364458278038 28140
                     |253827715945792003647537174378183292284083925805177
                   high----------------------------------------------------
10 3737   -3.57 E|.........0.......0000000000...0.......0......0000
 5 3722   -1.05 e|.....0.....00..0...0.0.....................................
 6 3723    -.90 C|.00.......0...00.....0.....................  ..........
 1 3689    -.65 b|........0       ...  ..  .. .........    ....  ..1.
 2 3698    -.57 a|....0....       ...  .   .. .........    ....  .1.1
11 3738    -.29 F|0..0..00...0.....................................
 7 3726     .29 B|.............................................1...
 4 3721     .70 c|.............. ......... .....1.1..........1......
 8 3729    1.24 f|.......................... ...   1.......1......111.
12 3740    1.88 d|.............. ........ 1........1...... ........
 3 3704    2.56 A|.................. ... 11. .1111111111111111111...1
                     |-----------------------------------------------low-
                     |22211111155431538114191742261192219954111121411977
                     |10097211054693244814658722130043286445521343824140
                     |73118895494579 0088 759 153364107322840878038 08177
                     |253827715        36   3  74 78 329      3925  5
```

Figure 7 – Table of extreme data

This table shows the extreme items. The green area (horizontal) line is indicated (No. 10, the order of questions), followed by the code issues, the value of -3.57 is the difficulty of questions. Green area (vertical) shows the people who surprised his response to this item.

7. Conclusion

Winsteps program can be a useful tool in forming a correct compilation of teaching tests; we can use it to distinguish the difficulty of each task. Graphical representations of difficulty are more clearly shows the distribution of items in the test. Too difficult a task or too easy task, so the test can be omitted and replaced by a more appropriate questions using extreme data, we determine what the issues were the respondents' answers surprising and unexpected, and vice versa, which corresponds to our surprise.

References

Burjan, V. (2005). Tvorba a využívanie školských testov v pedagogickej praxi. Bratislava: Metodicko-pedagogické centrum, s. 7, 33-34. ISBN 80-8052.228-6.

Lapitka, M. (1990). Tvorba a použitie didaktických testov. Bratislava: SPN, s. 31. ISBN 80-08-00782-6.